A51B-09

Non-supervized classification of ground-based radiometer retrievals (AERONET) in order to assess the distribution of aerosol volume size distributions and refractive indexes

L. Gross[1], R. Frouin[1], C. Pietras[2], K. Knobelspiesse[3] and G. Fargion[2]

1 Scripps Institution of Oceanography, University of California at San Diego, La Jolla, CA, USA
2 Sciences Applications International Corporation, SIMBIOS Project, NASA/GSFC, Greenbelt, MD, USA
3 Science Systems and Applications, Inc., SIMBIOS Project, NASA/GSFC, Greenbelt, MD, USA

Introduction

Typically, ocean color algorithms use aerosol mixture models to evaluate the atmospheric contribution to the signal (atmospheric correction) and then derive the oceanic content, indexed by chlorophyll-a concentration. The accuracy of ocean color retrievals from SeaWiFS, POLDER, OCTS, MODIS, MERIS, etc., relies on assumptions of the optical properties associated with each aerosol type. Gordon and Wang [1994] use nine Shettle and Fenn [1979] maritime and tropospheric aerosol models with a humidity variation of aerosol optical properties. A coastal aerosol model, composed of a maritime and tropospheric model mixture, was also used.

Shettle and Fenn [1979] developed their models using aerosol samples from the lower tropospheric aerosol samples, for which they derived the optical characteristics. In atmospheric correction, however, we are more interested in the optical behavior of the aerosols through the entire atmosphere. Comparisons of SeaWiFS-derived and in situ aerosol optical thickness values [Pietras et. al, 2001], on the other hand, have revealed a systematic underestimation of the Angstrom coefficient. This might be evidence that the reference models are not representative of actual conditions, although the discrepancy might also be due to the procedure of selecting models or radiometric calibration errors.

To provide answers to the above questions (i.e., representation of the models and origin of atmospheric correction errors), and ultimately improve atmospheric correction, one needs to analyze optical atmospheric data under varied aerosol conditions, over oceans around the world. The AERONET Program uses CIMEL radiometers, originally owned by the SIMBIOS Project, operating continuously at many island and coastal sites to monitor aerosol optical properties. The maturity of the CIMEL data processing procedures and inversion algorithms [Dubovik et al., 2000], allows us to make a global statistic on aerosol mixtures.

A non-supervised classification of the retrieved aerosol properties of the total atmospheric column, i.e. the volume size distribution function and the refractive index, may allow us to determine the natural distribution and more importantly to identify clusters in this distribution. These clusters may be used as new aerosols mixtures in radiative transfer algorithms. We show here a first attempt of classification, using a probabilistic self-organizing map (PRSOM) to approximate the distribution of the data, followed by a hierarchical clustering to identify geophysical conditions in the data base.

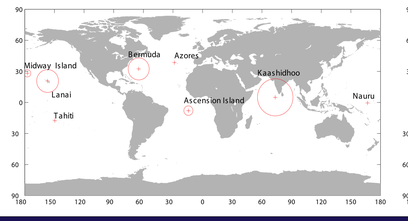
Data

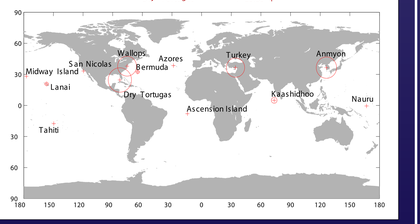
Figure 1 — The Data Set Distribution (%)

Nauru	: 0.37		
Tahiti	: 1.84		
Azores	: 2.21		
Midway Island	: 7.00		
Ascension Island	: 9.21		
Bermuda	: 21.36		
Lanai	: 22.10	79.37	
Kaashidhoo	: 35.91		

Figure 2 — The Data Set Distribution (%)

Nauru	: 0.06		
Tahiti	: 0.30		
Azores	: 0.36		
Midway Island	: 1.14		
Ascension Island	: 1.50		
San Nicolas	: 1.83		
Bermuda	: 3.47		
Lanai	: 3.59		
Kaashidhoo	: 5.83		
Turkey	: 18.61		
Wallops	: 19.15	81.93	
Anmyon	: 20.62		
Dry Tortugas	: 23.55		





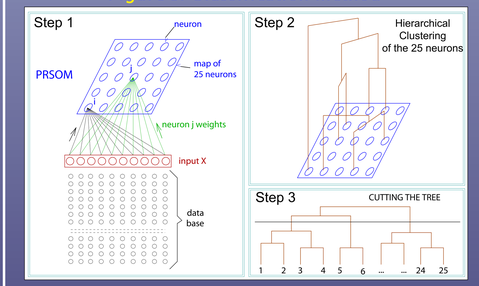
We classified particle volume size distributions (VDF) and corresponding refractive index (REF) retrievals from the AERONET inversion algorithm (level 1.5; screened for clouds) [Dubovik et al, 2000]. The algorithm computes VDF for particle radii ranging from 0.05 to 15 micrometers. The REF real and imaginary parts are computed at four wavelengths: 440, 670, 870 and 1020 nanometers. We kept AERONET retrievals whose VDF relative error was less that 7%, solar zenith angle was greater than 45 degrees and aerosol optical thickness was greater than 0.1, in order to minimize refractive index error. We normalized the VDF and approximated it by three log-normal distributions, each one parameterized by a magnitude A, a center [μ] and a standard deviation [ε]. One data vector (denoted X) is composed by 17 variables: the 9 variables of the VDF, the 4 variables of the REF real part and the 4 variables of the REF imaginary part. We performed two experiments, which correspond to different geographical ensembles. The first experiment (Exp. 1) is made using island sites exclusively, the second experiment (Exp. 2) merges island sites and coastal sites. Exp. 1 had 543 data points, among which 80% are on Bermuda, Lanai and Kaashidhoo (see Fig. 1). Exp. 2 had 3342 data points, among which 80% are on Turkey, Wallops, Anmyon and Dry Tortugas (coastal sites, see Fig. 2).

http://simbios.gsfc.nasa.gov

Contact: Lydwine Gross at lidwine@polaris.ucsd.edu

Method

Figure 3: Classification method



In order to reduce the number of data dimensions, we preprocessed the two data ensembles using principal component analysis (PCA) [Joliffe, 1986]. The number of variables for data vector X was reduced to 10. Our classification is performed using three steps (Fig. 3):

Step 1 - We summarize the information contained in the studied data set using a Probabilistic Self Organizing Map (PRSOM), which approximates the data distribution and leads to a definition of reference vectors (25 in this case). Self-Organizing Maps (SOM) are neural networks, which were first introduced by Kohonen [1994], for visualizing and clustering n-dimensional observations. SOM models have two layers, where the input layer is the number of neurons equal to the data space dimensions (here 10) and the topological map layer, which is a discrete lattice of neurons, connected to the neurons of the input layer by weighted synapses. Once the SOM is trained, each neuron, j, on the topological map represents a class of inputs, and is characterized by a reference vector composed of its weights. There is a topological relationship (neighborhood) between different classes.

Step 2 - Since we had to choose the number of PRSOM classes a priori, we make a hierarchical clustering of the reference vectors (with Ward dissimilarity) to control the number of statistically important clusters. We determine the natural cluster divisions of the data set by comparing the length of each link in the cluster tree with the lengths of neighboring links below it in the tree. If the length of a link differs from neighboring links, it indicates that there are dissimilarities between the objects at this level in the cluster tree. This link is said to be inconsistent with the links around it. In cluster analysis, inconsistent links can indicate the border of a natural division in a data set

Step 3 - Once the cluster tree (dendrogram) is cut, the reference vectors of each cluster are combined using a weighted mean which takes into account the number and the variance of the actual data attached to each reference. Finally we obtain N classes, each one containing a normalized particle volume size distribution and one refractive index.

Figure 4: The SOM Neurons Weights and their data cluster

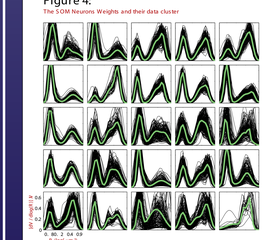


Figure 5:

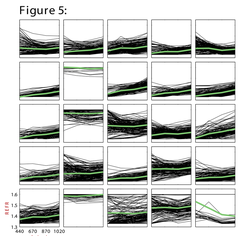
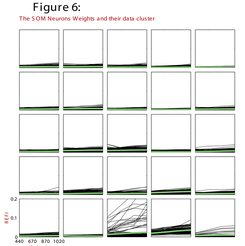


Figure 6: The SOM Neurons Weights and their data cluster



RESULTS of STEP 1: the reference vectors

Here we show here the results of Exp. 2. Each plot in Fig. 4, 5 and 6 stands for one neuron of the PRSOM (25 neurons total). In each plot we display the data gathered by the neuron (in black) and the corresponding reference vector which is used later in the classification (in green, the weights of the neuron). Note that for the PRSOM, a data point is a single particle volume size distribution function attached to one refractive index, but because these do not have the same units, they are presented on separate figures: Fig. 4 displays the particle volume size distribution function, Fig 5. displays spectra of REF real part, and Fig. 6 displays spectra of REF imaginary part.

Figure 7:

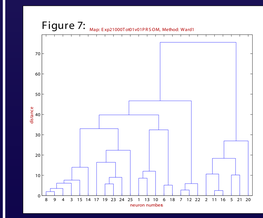
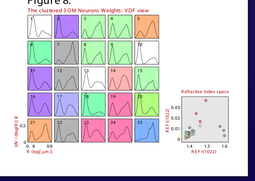


Figure 8: The clustered SOM Neurons Weights: VDF view



RESULTS of STEP 2: the hierarchical clustering of reference vectors

The reference vectors of Exp. 1 and Exp. 2 are clustered using Ward dissimilarity criteria (Fig. 7; Exp. 2 dendrogram). This method minimizes an incremental sum of squares; that is, the increase in the total within group sum of squares as a result of joining two groups. To initiate the linkage, a weight is assigned to each referent, which is the number of data points it represents. The computation of the link inconsistencies suggests to cut the dendrograms at the level of 11 connects. The result on Exp. 2 is displayed in Fig. 8. Note that a reference vector gathers the VDF and REF information, thus they are not classified separately. We have displayed the VDF information on a map. Each plot of the same color (standing for a neuron of PRSOM) belongs to the same class. On the map's right, the refractive index information is represented by the plot of the REF imaginary part at 1020 nm, versus the REF real part at 1020 nm of each neuron. The spots of same color belong to the same class. The colors match the VDF information colors.

Results

Figure 9:

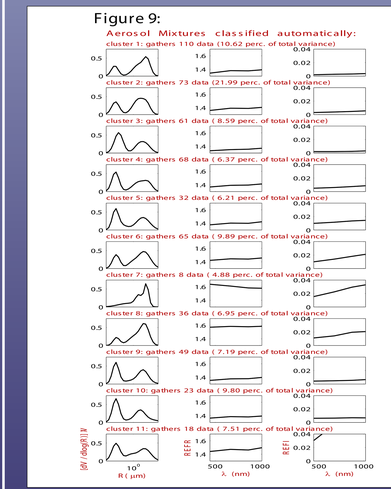
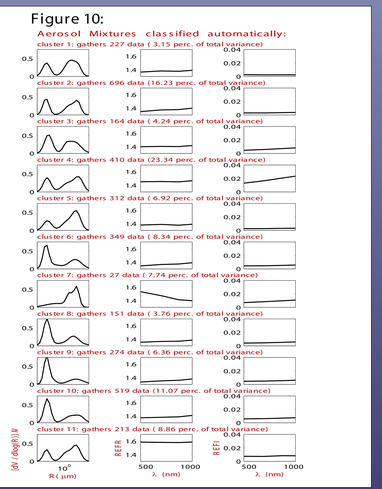


Figure 10:



A weighted mean of the reference vectors of each class allows us to obtain one VDF and one REF per class. The results on Exp. 1 and 2 are displayed in Fig. 9 and 10 respectively. We arbitrarily numbered the classes of each experiment. If we compute the euclidian distance matrix between the 11 final referents of Exp. 1 and 2, we find that 9 clusters in each experiment have an equivalent cluster in the other experiment, as is shown in Table 1. These stable clusters do not depend on the geographical distribution of the data base. Their expressions resulting from Exp. 2 should be statistically better as they have been built with much more data. Note that cluster 4 has strong absorption and cluster 11 has strong diffraction although they have a similar VDF. The two additional clusters of Exp. 2 (9 and 10) cannot be put aside as they represent a lot of data with low variance. These two classes must be specific to coastal sites as their fine mode is well developed.

Table 1: Cluster Matches

Exp. 2 cluster #		Exp. 1 cluster #
1	<-->	2
2	<-->	4
3	<-->	3
4	<-->	6
5	<-->	1
6	<-->	5
7	<-->	7
8	<-->	10
11	<-->	8

Conclusion

Two non-supervised classifications made on different geographical data sets allowed us to isolate nine stable clusters in the AERONET data base (Fig. 10 clusters 1 to 8 and cluster 11), and two clusters originating from coastal sites (Fig. 10 clusters 9 and 10). We now need to physically interpret these aerosol mixtures by comparing them with current models [Gordon and Wang, 1994], and more importantly, to determine their impact in terms of radiative transfer. The reference vectors of the clusters must be interpreted with respect to aerosol phase function and single scattering albedo using the Mie theory, so we could test if they improve satellite aerosol products. This will be performed on data from the SeaWiFS instrument.

Acknowledgments

This work is supported by the NASA SIMBIOS Project. We would like to thank A. Smirnoff, O. Dubovik and M. Wang for stimulating discussions. CIMEL data were gathered by the AERONET Program and the SIMBIOS Project. Principal Investigators for CIMEL sites included M. Miller (Brookhaven National Laboratories), C. McClain (NASA SIMBIOS Project), B. Holbren (NASA AERONET Program), R. Frouin (Scripps Institution of Oceanography, University of California at San Diego) and K. Voss (University of Miami).

References

Dubovik O., A. Smirnoff, B. N. Holben, M. D. King, Y. J. Kaufman, T. F. Eck and I. Slutsker 2000: Accuracy assessments of aerosol optical properties retrieved from Aerosol Robotic Network (AERONET) Sun and Sky radiance measurements, *Journal of Geophysical Research*, 105(D8), 9791-9806.

Gordon, H. R. and M. Wang 1994: Retrieval of water-leaving radiance and aerosol optical thickness over the oceans with SeaWiFS: a preliminary algorithm, *Applied Optics*, 33(3), 443-452.

I. T. Jolliffe, 1986: *Principal component analysis*, Springer-Verlag, New York.

T. Kohonen,1994: *Self-organizing map*, Springer-Verlag, Berlin.

Pietras, C., M. Miller, R. Frouin, T. Eck, B. Holben, and J. Marketon, 2001: "Calibration of Sun Photometers and Sky Radiance Sensors," In Fargion, G., R. Barnes, and C. McClain, In Situ Aerosol Optical Thickness Collected by the SIMBIOS Program (1997-2000); Protocols, Data QC and Analysis, *NASA Tech. Memo. 2001-209982*, NASA Goddard Space Flight Center, Greenbelt, Maryland, 11-21.

Shettle, E. P. and R. W. Fenn, 1979: Models for the aerosols of the lower atmosphere and the effect of the humidity variations on their optical properties, Air Force Geophys. Lab., TR--79--0214, Bedford, Mass.

Yacoub M., D. Frayssinet, F. Badran and S. Thiria, Classification based on Expert knowledge propagation using Probabilistic Self-Organizing Map: application to geophysics (2000), In: *Data analysis: scientific modeling and practical applications*, Springer-Verlag, Studies in classification, Data Analysis, and knowledge organization Series.